15-859: Information Theory and Applications in TCS **CMU: Spring 2013**

Lecture 2: Source coding, Conditional Entropy, Mutual Information

January 17, 2013

Lecturer: Venkatesan Guruswami *Scribe: David Witmer*

1 The Shannon code

Consider a random variable X taking on values $a_1, \ldots, a_n$ with probabilities $p_1, \ldots, p_n$. We would like to encode values of X so that the expected number of bits used is small. Let $\ell_1, \ldots, \ell_n$ be the number of bits used to encode $a_1, \ldots, a_n$. Recall Kraft's Inequality from the previous lecture: A prefix-free code for X exists if and only if

$$\sum_{i=1}^{n} 2^{-\ell_i} \leq 1.$$

Furthermore, we can construct such a code from ℓ_i's satisfying this inequality, so we need to find good values for the ℓ_i's. In the Shannon code (sometimes called the Shannon-Fano code), we set

$$\ell_i = \lceil \log \frac{1}{p_i} \rceil.$$

The following calculation shows that these ℓ_i's satisfy Kraft's Inequality:

$$\sum_{i=1}^{n} 2^{-\ell_i} = \sum_{i=1}^{n} 2^{-\lceil \log \frac{1}{p_i} \rceil} \leq \sum_{i=1}^{n} 2^{-\log \frac{1}{p_i}} = \sum_{i=1}^{n} p_i = 1.$$

The expected length of the encoding is

$$\sum_{i=1}^{n} p_i \lceil \log \frac{1}{p_i} \rceil.$$

This is lower bounded by $H(X)$ and upper bounded by $H(X) + 1$:

$$\sum_{i=1}^{n} p_i \lceil \log \frac{1}{p_i} \rceil \geq \sum_{i=1}^{n} p_i \log \frac{1}{p_i} = H(X)$$

$$\sum_{i=1}^{n} p_i \lceil \log \frac{1}{p_i} \rceil < \sum_{i=1}^{n} p_i (\log \frac{1}{p_i} + 1) = H(X) + \sum_{i=1}^{n} p_i = H(X) + 1.$$

In some cases,the Shannon code does not perform optimally. Consider a Bernoulli random variable X with parameter 0.0001. An optimal encoding requires only one bit to encode the value of X. The Shannon code would encode 0 by 1 bit and encode 1 by $\log 10^4$ bits. This is good on average but bad in the worst case.

We can also compare the Shannon code to the Huffman code. The Huffman code always has shorter expected length, but there are examples for which a single value is encoded with more bits by a Huffman code than it is by a Shannon code. Consider a random variable X that takes values a, b, c, and d with probabilities 1/3, 1/3, 1/4, and 1/12, respectively. A Shannon code would encode a, b, c, and d with 2, 2, 2, and 4 bits, respectively. On the other hand, there is an optimal Huffman code encoding a, b, c, and d with 1, 2, 3, and 3 bits respectively. Note that c is encoded with more bits in the Huffman code than it is in the Shannon code, but the Huffman code has shorter expected length. Also note that the optimal code is not unique: We could also encode all values with 2 bits to get the same expected length.

2 Entropy lower bounds the expected length of encoding

Theorem 2.1 *The expected length of encoding by a prefix-free code of a random variable X is at least $H(X)$.*

Proof. The expected length is $\sum_{i=1}^{n} p_i \ell_i$. We will show that $H(X) - \sum_{i=1}^{n} p_i \ell_i \leq 0$. First, observe that

$$\begin{aligned} H(X) - \sum_{i=1}^{n} p_i \ell_i &= \sum_{i=1}^{n} p_i \log \frac{1}{p_i} - \sum_{i=1}^{n} p_i \ell_i \\ &= \sum_{i=1}^{n} p_i \log \frac{1}{p 2^{\ell_i}}. \end{aligned}$$

Define a random variable Y such that $Y = \frac{1}{p_i 2^{\ell_i}}$ with probability p_i. Then we have that

$$\begin{aligned} H(X) - \sum_{i=1}^{n} p_i \ell_i &= \mathbb{E}[\log Y] \\ &\leq \log \mathbb{E}[Y] \quad \text{by Jensen's Inequality} \\ &= \log \left(\sum_{i=1}^{n} p_i \frac{1}{p_i 2^{\ell_i}} \right) \\ &= \log \left(\sum_{i=1}^{n} 2^{-\ell_i} \right) \\ &\leq \log 1 \quad \text{by Kraft's Inequality} \\ &= 0. \end{aligned}$$

■

Since Kraft's Inequality also holds for non-prefix-free codes, $H(X)$ lower bounds the expected length of non-prefix-free codes as well.

3 Improving the Shannon code and the Fundamental Source Coding Theorem

As was shown in the first section, the Shannon code may have expected length greater than the optimal. By amortizing this loss over many symbols, we can approach an expected length equal to the entropy lower bound. We will assume that we have a source outputting a sequence of i.i.d. draws from a random variable X. Note that this assumption does not hold in many cases, e.g., the English language. We then have m independent draws $X_1 X_2 \ldots X_m$ from X. We also have the following fact:

Fact 3.1 $H(X_1 X_2 \ldots H_m) = mH(X)$.

We can then construct a Shannon code for the whole string $X_1 X_2 \ldots X_m$ of m values with expected length at most $mH(X) + 1$. The average number of bits per symbol is then at most $H(X) + \frac{1}{m}$. By delaying and sending symbols in larger chunks, we can get better amortized cost and spread out the loss of the Shannon code over many symbols. This proves the Fundamental Source Coding Theorem, also called the Noiseless Coding Theorem.

Theorem 3.2 (Fundamental Source Coding Theorem) *For all $\varepsilon > 0$ there exists n_0 such that for all $n \geq n_0$, given n i.i.d. samples $X_1 X_2 \ldots X_n$ from a random variable X, it is possible to communicate at most $H(X) + \varepsilon$ bits per sample on average to reveal $X_1 X_2 \ldots X_n$ to the other party.*

This theorem also holds with high probability, not just on average.

4 Joint and conditional entropy

Definition 4.1 (Joint entropy) *Let X and Y be two possibly correlated random variables. The joint entropy of X and Y, denoted $H(X,Y)$, is*

$$H(X,Y) = \sum_{x,y} p(x,y) \log \frac{1}{p(x,y)},$$

where $p(x,y)$ is defined to be $\Pr(X = x \wedge Y = y)$.

If X and Y are independent, $p(x,y) = p(x)p(y)$ and

$$H(X,Y) = \sum_{x,y} p(x)p(y) \left(\log \frac{1}{p(x)} + \log \frac{1}{p(y)} \right) = H(X) + H(Y).$$

In general,

$$H(X,Y) = \sum_{x,y} p(x,y) \log \frac{1}{p(x)p(y|x)},$$

where $p(y|x) = \Pr(Y = y|X = x)$.

We can then do the following calculation:

$$\begin{aligned}
H(X,Y) &= \sum_{x,y} p(x,y) \log \frac{1}{p(x)p(y|x)} \\
&= \sum_{x,y} p(x,y) \log \frac{1}{p(x)} + \sum_{x,y} p(x,y) \log \frac{1}{p(y|x)} \\
&= \sum_{x} p(x) \log \frac{1}{p(x)} + \sum_{x} p(x) \sum_{y} p(y|x) \log \frac{1}{p(y|x)} \\
&= H(X) + \sum_{x} p(x) H(Y|X = x) \\
&= H(X) + \mathbb{E}_x[H(Y|X = x)].
\end{aligned}$$

This motivates the definition of conditional entropy:

Definition 4.2 (Conditional entropy) *The conditional entropy of Y given X is*

$$H(Y|X) = \mathbb{E}_x[H(Y|X = x)].$$

Our calculation then shows this lemma:

Lemma 4.3 $H(X,Y) = H(X) + H(Y|X)$.

Intuitively, this says that how surprised we are by drawing from the joint distribution of X and Y is how surprised we are by X plus how surprised we are by Y given that we know X already.

Note that if X and Y are independent, $H(Y|X) = H(Y)$ and $H(X,Y) = H(X) + H(Y)$.

Recall the chain rule for probability: $p(x,y) = p(x)p(y|x)$, or, more generally,

$$p(x_1, \ldots, x_n) = p(x_1)p(x_2|x_1) \ldots p(x_n|x_1, \ldots, x_n).$$

There is a similar chain rule for entropy:

Theorem 4.4 (Chain rule) *For random variables X, Y, and Z,*

$$H(X,Y,Z) = H(X) + H(Y|X) + H(Z|X,Y).$$

For n random variables $X_1, \ldots, X_n$,

$$H(X_1, X_2, \ldots, X_n) = H(X_1) + H(X_2|X_1) + \cdots + H(X_n|X_1, X_2, \ldots, X_{n-1}).$$

The log in the definition of entropy changes the multiplication in the probability chain rule to addition. Also, the order of the random variables does not matter. For example, it also holds that

$$H(X,Y) = H(Y) + H(X|Y).$$

Note that $H(X|X) = 0$.

Example 4.5 *Let X be a random variable that is uniform on $\{0,1,2,3\}$. Let $Y = X \mod 2$.*

Clearly, $H(X) = 2$.

$H(Y) = 1$ since Y is uniform on $\{0,1\}$.

$H(X|Y) = 1$ because knowing Y tells us if X is odd or even.

$H(Y|X) = 0$ since knowing X tell us the exact value of Y.

$H(X,Y) = 2$ because X tells us everything about X and Y.

Intuitively, it seems like conditioning should never increase entropy: knowing more should never increase our surprise. This is indeed the case:

Lemma 4.6 (Conditioning cannot increase entropy) $H(Y|X) \leq H(Y)$.

Proof. The proof is similar to the above proof that the entropy lower bounds the expected length of a code. First, we have that

$$\begin{aligned} H(Y|X) - H(Y) &= H(X,Y) - H(X) - H(Y) \\ &= \sum_{x,y} p(x,y) \log \frac{1}{p(x,y)} - \sum_x p(x) \log \frac{1}{p(x)} - \sum_y p(y) \log \frac{1}{p(y)}. \end{aligned}$$

Since $p(x) = \sum_y p(x,y)$ and $p(y) = \sum_x p(x,y)$,

$$H(Y|X) - H(Y) = \sum_{x,y} p(x,y) \log \frac{p(x)p(y)}{p(x,y)}.$$

We now define Z to be a random variable taking value $\frac{p(x)p(y)}{p(x,y)}$ with probability $p(x,y)$, so

$$\begin{aligned} H(Y|X) - H(Y) &= \mathbb{E}_{x,y}[\log Z] \\ &\leq \log \mathbb{E}[Z] \quad \text{by Jensen's Inequality} \\ &= \log \left(\sum_{x,y} p(x,y) \frac{p(x)p(y)}{p(x,y)} \right) \\ &= \log \left(\left(\sum_x p(x) \right) \left(\sum_y p(y) \right) \right) \\ &= \log 1 \\ &= 0. \end{aligned}$$

■

As a corollary, we have a statement similar to the union bound:

Corollary 4.7 *For random variables X and Y,*

$$H(X,Y) \leq H(X) + H(Y).$$

More generally, for random variables $X_1, \ldots, X_n$,

$$H(X_1, \ldots, X_n) \leq \sum_{i=1}^{n} H(X_i).$$

Exercise 4.8 *For a random variable X with support size n, we can think of entropy as a function from $[0,1]^n$ to $\mathbb{R}_{\geq 0}$. If X takes on n different values with probabilities $p_1, \cdots, p_n$, then for $\mathbf{p} = (p_1, \cdots, p_n)$, $H(\mathbf{p}) = \sum_{i=1}^{n} p_i \log \frac{1}{p_i}$. Show that $H(\mathbf{p})$ is a concave function, i.e., show*

$$H(\lambda \mathbf{p} + (1-\lambda)\mathbf{q}) \geq \lambda H(\mathbf{p}) + (1-\lambda) H(\mathbf{q})$$

for all $\lambda \in [0,1]$, $\mathbf{p}, \mathbf{q} \in [0,1]^n$.

5 Mutual information

Definition 5.1 (Mutual information) *The mutual information between random variables X and Y, denoted $I(X;Y)$, is*

$$I(X;Y) = H(X) - H(X|Y).$$

Intuitively, mutual information is the reduction in the uncertainty of X that comes from knowing Y.

We can write $I(X;Y)$ in several other equivalent ways:

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= H(X) - (H(X,Y) - H(Y)) \\ &= H(X) + H(Y) - H(X,Y) \\ &= H(Y) - H(Y|X) \\ &= I(X;Y) \end{aligned}$$

Note that $I(X;Y) = I(Y;X)$.

The next lemma follows from the fact that conditioning cannot increase entropy.

Lemma 5.2 $I(X;Y) \geq 0$.

Also, if X and Y are independent, $I(X;Y) = 0$.

Example 5.3 *Consider X and Y as defined in Example 4.5. Then*

$$I(X;Y) = H(X) - H(X|Y) = 2 - 1 = 1.$$

Example 5.4 *Let the Z_i's be i.i.d. random variables that are uniform over $\{0,1\}$. Let $X = Z_1 Z_2 Z_3 Z_4 Z_5$ and $Y = Z_4 Z_5 Z_6 Z_7$. Then $I(X;Y) = 2$ since X and Y have 2 bits in common.*

We show the relationship between entropy, joint entropy, conditional entropy, and mutual information for two random variables X and Y in Figure 5.1.

We can also define a conditional version of mutual information.

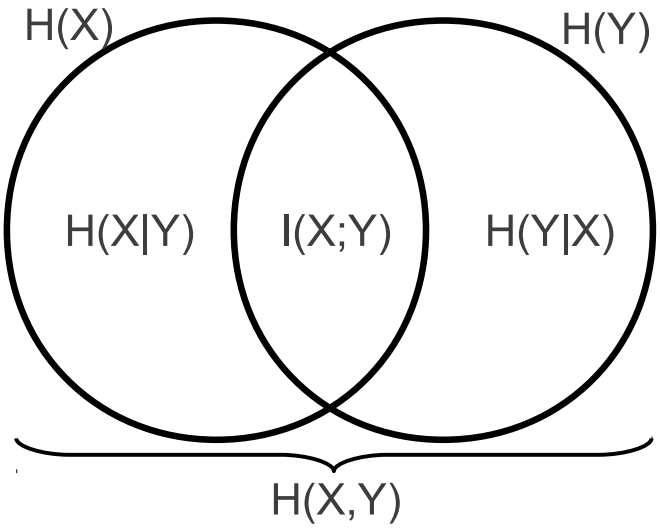

Figure 5.1: Relationship between entropy, joint entropy, conditional entropy, and mutual information for two random variables.

Definition 5.5 (Conditional mutual information) *The conditional mutual information between X and Y given Z is*

$$\begin{aligned} I(X,Y;Z) &= H(X|Z) - H(X|Y,Z) \\ &= H(Y|Z) - H(Y|X,Z). \end{aligned}$$

Exercise 5.6 *Prove the chain rule for mutual information:*

$$I(X_1, X_2, \ldots, X_n; Y) = \sum_{i=1}^{n} I(X_i; Y | X_1, X_2, \ldots, X_{i-1}).$$

Note that the order of the X_i's does not matter.

Exercise 5.7 *Is it always true that $I(X;Y|Z) \leq I(X;Y)$? Give a proof or a counterexample.*